Knowledge Discovery

Profiling microblog authors using concreteness and sentiment

Overview

Problem

The PAN 2016 author profiling task is a supervised classification problem on cross-genre documents (tweets, blog and social media posts). The task presents two separate classification problems: gender classification and age group classification. The latter is a multi-class (18-24, 25-34, 35-49, 50-64 and 65+) classification problem. The classification problem can be described as follows: An author profile in the context of the task is defined as an author's gender and age group. Given a set of documents with author profiles known, learn to identify the author's profile of documents of unknown authorship.

Approach

Our As mentioned in the problem description, we consider the problem a supervised classification problem. We pre-process each document, extracting features and thus vectorising the input. Once all the features a extracted, we train a random forest model. The random forest implementation we use is provided by the class *RandomForest* from the machine learning framework WEKA [Hall2009]. We did not tune any parameters, but used WEKA's default settings.

Features

Concreteness

A number of features are based on the concreteness of words within tweets. The base of this features is a dataset assembled by Brysbaert et al. with the help of Amazon Mechanical Turk [1].

The dataset comprises over 37 thousand words known to a large share of the English speaking population. Concreteness is defined in this context, whether a word refers to a perceptible entity, driven by the intuition that concrete words are easier to remember and to process than words that refer to abstract concepts.

Concreteness has been studied in a variety of scenarios, containing the tendency to use words with varying degree of concreteness depending on age and gender [2].

Our set of concreteness features consists of nine individual numeric features, based on three different scores being computed on a per word basis:

1. Mean concreteness: The score reflects the concreteness of the words within a tweet. Concreteness thereby ranges from 5 to 1.
2. Standard deviation concreteness: This score encodes how strong the individual annotators agreed on the concreteness score. For words were all raters agreed, the score will be low.
3. Percent known: This score represent the percentage of all raters, who indicated that they know the word. This score ranges from 0.85 to 1.

In order to arrive at features at tweet level, all word based scores are aggregated. Therefore the minimum, the maximum and the arithmetic mean are computed for each of the three types of scores.

Sentiment

We use the well known sentiment library called SentiWordNet [3], which provides a linear score between -1 and 1 to specify the polarities of words depending on their context. We extract the score of the most used context for each token in SentiWordNet. We then model the polarity as four numeric features, encoding the maximum, minimum and average polarity as well as the standard deviation of polarity. These features represent the polarity distribution of a tweet seen as a bag of words.

WordNet Domains

To encode the main topics of a tweet in a concise way we used the publicly available WordNet Domains corpus [4], an augmented version of WordNet [5] assigning words to about 200 hierarchically ordered domains.

Our algorithm creates a set of domains for a given short snippet of text. All domains of all words are combined while keeping track of a weight. The weight reflects how ambiguous the domain mappings are, thus words with many domains will yield lower weights.

Finally, the hierarchy of the domains is exploited, where each sub-domain distributed a share of its current weight to its parent. The ranked list of domains is finally pruned. All domains with a lower weight than half of the weight of the top ranked domain will be removed. On average a short snippet of text will yield a set of 1 to 5 domains.

In order to convert the set of domains into features we created a binary feature for each domain. If a tweet is associated with a certain domain, the corresponding feature will be set to true.

Token Length

Users familiar with micro blogging or texting are used to the 140 character limit. As a consequence, we expect more frequent usage of abbreviations and acronyms from such users. We encode the mean token length and the median token length.

Results

For memory limitations on the validation system we deactivated the WordNet domains feature group. Here are the evaluation results obtained from TIRA.

Dataset	Gender	Age Class	Both
dataset2-english-2016-05-07	0.5769	0.3205	0.1410
dataset1-english-2016-03-08	0.0201	0.0086	0.0057

While the results on 'dataset2' are where we expected them to be, the results on the results on 'dataset1' are extremely low. We cannot comment on this yet, as we have no further details on how the test sets look like.

Publications

O. Pimas, A. Rexha, M. Kröll, and R. Kern, "Profiling microblog authors using concreteness and sentiment - Know-Center at PAN 2016 author profiling", CLEF2016 Work. Notes, 2016.

References

[1] Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46(3), 904–911. article.

[2] Calais, L. L., Lima-Gregio, A. M., Arantes, P., Gil, D., & Borges, A. C. L. de C. (2012). A concreteness judgment of words. Jornal Da Sociedade Brasileira de Fonoaudiologia, 24(3), 262–268. article.

[3] Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In 5th Language Resources and Evaluation Conference (LREC 2006) (pp. 417–422). inproceedings.

[4] Magnini, B., & Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. In LREC. inproceedings.

[5] Miller, G. a. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41. http://doi.org/10.1145/219717.219748

[6] Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, London, England. http://doi.org/10.1139/h11-025

[7] Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. (2014). Improving the reproducibility of PAN's shared tasks: Plagiarism detection, author identification, and author profiling. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8685 LNCS, pp. 268–299). http://doi.org/10.1007/978-3-319-11382-1_22

[8] Gollub, T., Stein, B., Burrows, S., & Hoppe, D. (2012). TIRA: Configuring, executing, and disseminating information retrieval experiments. In Proceedings - International Workshop on Database and Expert Systems Applications, DEXA (pp. 151–155). http://doi.org/10.1109/DEXA.2012.55

Oliver Pimas
Know-Center GmbH
opimas@know-center.at

Acknowledgements: The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.